Grenze International Journal of Engineering and Technology, January Issue

Student Performance Prediction Model using Machine Learning Algorithms

Pranoy Das[1] and Shanu K. Rakesh[2]
[1-2]Chouksey Engineering College, Bilaspur, Chhattisgarh, India
Email: pranoydas1999@gmail.com, shanu.kuttan28@gmail.com

***Abstract*—Predicting student performance has become an increasingly vital aspect of educational research, providing educators with valuable insights for personalized learning interventions. This model is a comprehensive student performance prediction model using various machine learning algorithms like Logistic Regression, Naïve Bayes Classifier, SVM and K-Nearest Neighbors. The developed model not only provides valuable insights for educators to identify at-risk students but also enables the implementation of targeted interventions to improve student outcomes. Furthermore, the model's flexibility allows for customization according to different educational contexts, making it a valuable tool for educational institutions striving to enhance student success and retention rates. When compared to the results of the most recent research done on this subject, the proposed model provides a better result in terms of accuracy of the model.**

***Keywords* - Student performance prediction, Machine learning, Academic support.**

I. INTRODUCTION

In recent years, predicting student performance has become increasingly important in educational research. The ability to accurately forecast student outcomes allow educators to intervene early, providing targeted support and interventions to improve student success rates. Traditional methods of prediction have often relied on demographic information and historical grades. However, with the advancement of machine learning techniques and the availability of richer datasets, there is an opportunity to develop more accurate and comprehensive prediction models.

I got the motivation for this from my years working as a faculty in multiple renowned organizations. And as a faculty, I have struggled to keep records of such a mass of students that have their academic capabilities and understanding distributed over such a wide spectrum of possibilities.

And this only makes the work of predicting a specific student's performance in the upcoming examinations even harder. And this became my motivation, to develop a system, that uses extremely basic features about a student to identify and detect patterns in their academic performance and based on those patterns make accurate predictions about their performance in the upcoming examinations.

This paper presents a student performance prediction model using machine learning techniques. The model utilizes a dataset comprising 6000 student records, incorporating a wide range of features such as speaking skills, writing skills, class behavior, and examination percentages. By leveraging machine learning algorithms, this model can accurately predict student performance and identify at-risk students early.

Grenze ID: 01.GIJET.11.1.321

The developed model will be extremely beneficial for teachers who would be keen to know about the strong and weak points of their dear students, talk about the same with them as well as their parents, make communication based on factual data, to predict a trajectory on which the student is based on his/her current performance and make the most appropriate plan of action for that particular student to maximize his/her academic performance.
This model works best for the primary and middle school students, as the features used in the dataset are centered around the students who are studying in these grades, where it is extremely crucial to provide a proper guidance to them in their early years of education.

II. LITERATURE SURVEY

G. M. Rao and Prof. K. Kiran Kumar – (2021) [1] - Automatic Student performance prediction is a crucial job due to the large volume of data in educational databases. This job is being addressed by educational data mining (EDM). EDM develop methods for discovering data that is derived from educational environment. These methods are used for understanding student and their learning environment. Several well-known classification algorithms are applied in this domain existing models a student performance prediction model based on supervised learning decision tree classifier. In addition, an ensemble method is applied to improve the performance of the classifier. Ensemble methods approach is designed to solve classification, prediction problem. we propose a method for predicting final grades of students by a Recurrent Neural Network (RNN) from the log data stored in the educational systems. We applied this method to the log data from 108 students and examined the accuracy of prediction. The proposed model achieved an accuracy of 84.3%. With larger dataset records and features, an RNN can achieve higher accuracy and will outperform other machine learning algorithm.
C. A. Palacios, José A. Reyes-Suárez, Lorena A. Bearzotti, Víctor Leiva and Carolina Marchant – (2021) [2] - Data mining is employed to extract useful information and to detect patterns from often large data sets, closely related to knowledge discovery in databases and data science. In this investigation, we formulate models based on machine learning algorithms to extract relevant information predicting student retention at various levels, using higher education data and specifying the relevant variables involved in the modeling. Then, we utilize this information to help the process of knowledge discovery. We predict student retention at each of three levels during their first, second, and third years of study, obtaining models with an accuracy that exceeds 80% in all scenarios. These models allow us to adequately predict the level when dropout occurs.
Harikumar Pallathadka et al. – (2021) [3] - In today's competitive world, it is critical for an institute to forecast student performance, classify individuals based on their talents, and attempt to enhance their performance in future tests. Students should be advised well in advance to concentrate their efforts in a specific area in order to improve their academic achievement. Algorithms are evaluated based on characteristics such as accuracy and error rate. SVM is the most accurate technique for classifying a data set of student performance with an accuracy of about 87%.
Leila Ismail, Huned Materwala and Alain Hennebelle – (2021) [4] - Machine learning for education is an emerging discipline where a model is developed based on training data to make predictions on students' performance. The main aim is to identify students who would have difficulty in their learning and to take precautionary measures to help them. In this paper, we conduct a comparative analysis of the most used machine learning classification models in the literature. The experimental results reveal that for a dataset having fewer observations, SVM linear, SVM polynomial, and NB out performs the other models under study, whereas for a dataset having a large number of observations, DT and RF outperforms the other models under study. However, for generalizing these results, more extensive experiments are required on several balanced and imbalanced datasets.
Nitin Ramrao Yadav and Sonal Sachin Deshmukh – (2022) [5] - Machine learning is an aspect of artificial intelligence in which a computing system can able to learn from data and make conclusions. Student performance assessment is an important measurement metrics in education which affects the university accreditation. This review shows the intended results of Learning Analytics, Prediction Analytics and different Machine Learning algorithms and Data Mining techniques to improve the student outcome. This study results that machine learning technique is most used approach to enhance student performance and helps teachers to predict student achievements. The most widely used Machine Learning algorithms to enhance student performance at entry level and during academic year are Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayes (NB), Linear Regression (LR) and Decision Tree (DT).
Ziling Chen, Gang Cen, Ying Wei and Zifei Li – (2023) [6] - Predicting student performance is crucial for improving students' future academic achievements. Within student groups, common characteristics can reveal

trends in overall student learning. Most studies tend to focus on the common characteristics of students but ignore their individual characteristics. The results show that the implementation of the model on real student data can effectively predict student performance with an accuracy of 93.1% and an F1-score of 90.45%. With the model's predicted student performance, educators can provide individualized support and assistance to each student.

Esmael Ahmed – (2024) [7] - Modern learning institutions face challenges in analyzing performance, providing high-quality education, formulating strategies for evaluating students' performance, and identifying future needs. E-learning is a rapidly growing and advanced form of education, where students enroll in online courses. The study found that the SVM algorithm had the best prediction results after parameter adjustment, with a 96% accuracy rate. The researchers have examined the functions of the Support Vector Machine, Decision Tree, naive Bayes, and KNN classifiers. The outcomes of parameter adjustment greatly increased the accuracy of the four prediction models. Naïve Bayes model's prediction accuracy is the lowest when compared to other prediction methods, as it assumes a strong independent relationship between features

III. PROPOSED WORK

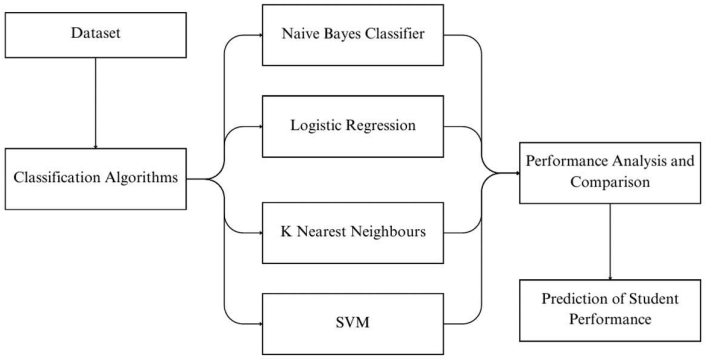

Figure 1. Proposed Work

The objective of this model is to predict the performance of student utilizing various machine learning algorithms.

Figure 1 shows the proposed work. Where the first step is dataset, select the dataset for analysis and comparison. The dataset may include academic records, attendance, and behavioral data. Second step is the classification algorithms where different algorithms are used for comparison. In the next step performance is analysed and algorithms are compared to find out the best algorithm based on multiple evaluation metrics such as accuracy, precision, recall and F1-score. At last, the proposed model will predict the student performance and categorize it in one of the three parameters - Score <= 50%, 50% <= Score <= 80% and Score > 80%.

The dataset comprises 6000 student records custom created from scratch. The dataset includes the following attributes:

- Student name
- Grade of the student
- Speaking skills: Assessed based on oral presentations and class participation.
- Writing skills: Assessed based on essays, homework assignments, and written exams.
- Spelling mistakes: Assessed based their copies and answer writing skills and word selections.
- Class participation: Assessed based on the how active the student is during the session's duration, like answering questions, asking doubts, peer communication, etc.
- Class behavior: Assessed based on attendance, participation, and disciplinary incidents.
- Assignment Submission: Assessed based on the number of assignments or homework completed by the student and the total number of assignments or homework provided to the student.
- Examination percentages: Grades from previous exams and assessments. These percentages are collected over the number of major examinations that take place in their respective school. So, the number of these Examination percentages might vary based on the number of students and the schools from where they belong.

- Average Examination percentage: This is the average of all the percentages that the student has gotten in the school examinations.
- Student performance: This is the output feature which would predict as in which group does the student belong. There are 3 values that are possible for this group: Score <= 50%, 50% <= Score <= 80% and Score > 80%.

IV. ALGORITHMS

A. Logistic Regression.

Logistic regression is a widely-used statistical and machine learning technique for binary classification problems, where the goal is to predict one of two possible outcomes. It works by modeling the probability of a given input belonging to a certain class. It predicts probabilities for two possible outcomes, often denoted as 0 or 1.
The core of logistic regression is the sigmoid function, which converts the linear combination of input features into a probability value between 0 and 1.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where z is the output of the logistic regression.

B. Naïve Bayes Classifier

The Naive Bayes classifier is a probabilistic machine learning algorithm based on Bayes' Theorem, used for classification tasks. It assumes that the features used for classification are conditionally independent given the class label, which simplifies the computation and makes it "naive".
The classifier applies Bayes' Theorem to calculate the posterior probability of a class given the features

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$

Where P(C|X) is posterior probability of class C given features X, P(X|C) is the likelihood of features X given class C, P(C) is the prior probability of class C, and P(X) is the marginal probability of features X.

C. Support Vector Machines (SVM)

Support Vector Machines (SVMs) are a supervised learning algorithm used for classification and regression tasks. The basic idea behind SVMs is to find the hyperplane that best separates the data points of different classes in the feature space. SVMs are particularly effective in high-dimensional spaces and are capable of capturing complex relationships in the data through the use of kernel functions.
The SVM algorithm operates as follows:

- Given a set of labeled training data, the algorithm finds the hyperplane that maximizes the margin between the classes.
- If the data is not linearly separable, the algorithm maps the input features into a higher-dimensional space using a kernel function.
- The algorithm then finds the optimal hyperplane in the higher dimensional space that best separates the data.

SVMs are known for their ability to handle high-dimensional data, their effectiveness in capturing complex relationships, and their resistance to overfitting.

D. K Nearest Neighbors

The k-Nearest Neighbors (k-NN) algorithm is a simple, instance-based machine learning technique used for classification and regression tasks. It classifies a data point based on the majority class among its k nearest neighbors or predicts a value based on the average of its nearest neighbors. KNN does not require a training phase in the traditional sense. Instead, it stores the training examples and performs classification or regression at prediction time by comparing new data points to these stored examples. To determine "nearest" neighbors, k-NN uses a distance metric, typically Euclidean distance, but other metrics like Manhattan or Minkowski can also be used.

$$Distance = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where x_i and y_i are feature values of the two points being compared.

V. METHODOLOGY

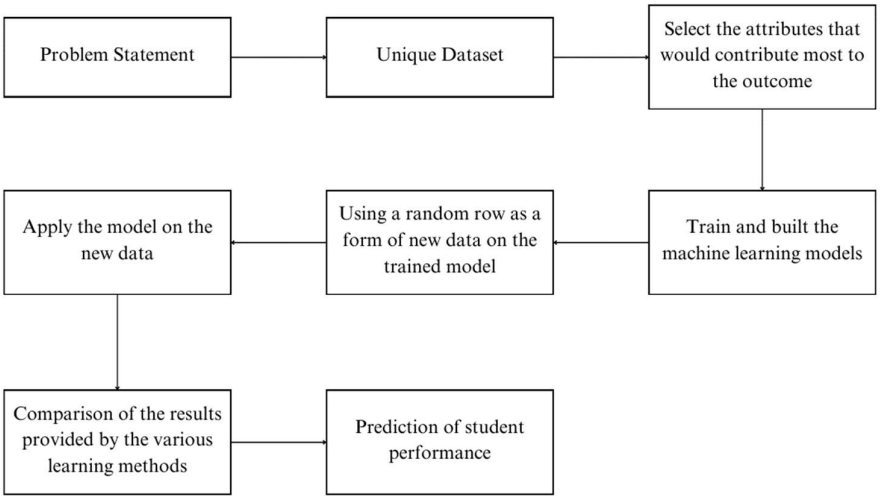

Figure 2. Design and Approach

Figure 2 shows the project plan where first step is to identify the problem statement, where the problem is identified to accurately predict the performance of a student. The next step is the selection of the appropriate dataset. For this purpose, I have used my own unique dataset, based on unique features of its own, which I used when I was a faculty and I had to know whether one of my students will be performing great in the examinations or not. This dataset already contained the most vital set of attributes that will contribute to the student performance one way or the another. The next step is planning the problem-solving strategy, for this using six classifiers, they are is SVM i.e. support vector machine, Navie Bayes, KNN and logistic Regression. The next step is building a training model for prediction. In this step, the model is trained by using different machine learning algorithms and the results were obtained by using different performance metrics. The next step is applying the model on a new data, which is nothing but a random data picked from the main dataset. The next step is comparison of various machine learning algorithm based on evaluation metrics. In the last step it will compare all the classifiers and find out about the academic performance of a particular student.

VI. RESULTS AND DISCUSSION

This model works for predicting the performance of a student by employing machine learning algorithm over a dataset comprising 6000 student records and incorporating features such as speaking skills, writing skills, class behavior, and examination percentages, the model provides an accurate and efficient method for predicting student performance based on which algorithm is applied for making predictions on the target dataset. The results demonstrate the potential of SVM and KNN as a valuable tool for educators seeking to improve student outcomes with an accuracy of about 98% and 97% respectively. On evaluating the performance of the proposed model on the evaluation metrics such as confusion metrics, accuracy, precision, recall and f1-score, the Table I shows the comparative visualization of the performance of all the employed algorithms.

TABLE I. OUTCOMES OF EVALUATION METRICS FOR ALL THE EMPLOYED ALGORITHMS

Algorithm	Confusion Matrix	Accuracy	Precision	Recall	f1-Score
Logistic Regression	$\begin{bmatrix} [651 & 143 & 0] \\ [150 & 789 & 12] \\ [0 & 24 & 31] \end{bmatrix}$	0.82	0.79	0.74	0.76

SVM	$\begin{bmatrix} [794 & 0 & 0] \\ [11 & 940 & 0] \\ [0 & 16 & 39] \end{bmatrix}$	0.98	0.99	0.9	0.94
KNN	$\begin{bmatrix} [774 & 20 & 0] \\ [27 & 920 & 4] \\ [0 & 0 & 55] \end{bmatrix}$	0.97	0.96	0.98	0.97
Naive Bayes Classifier	$\begin{bmatrix} [792 & 2 & 0] \\ [152 & 775 & 24] \\ [0 & 1 & 54] \end{bmatrix}$	0.9	0.84	0.93	0.87

Based on the above table, below is the comparative visualization of the performance of all the employed algorithms

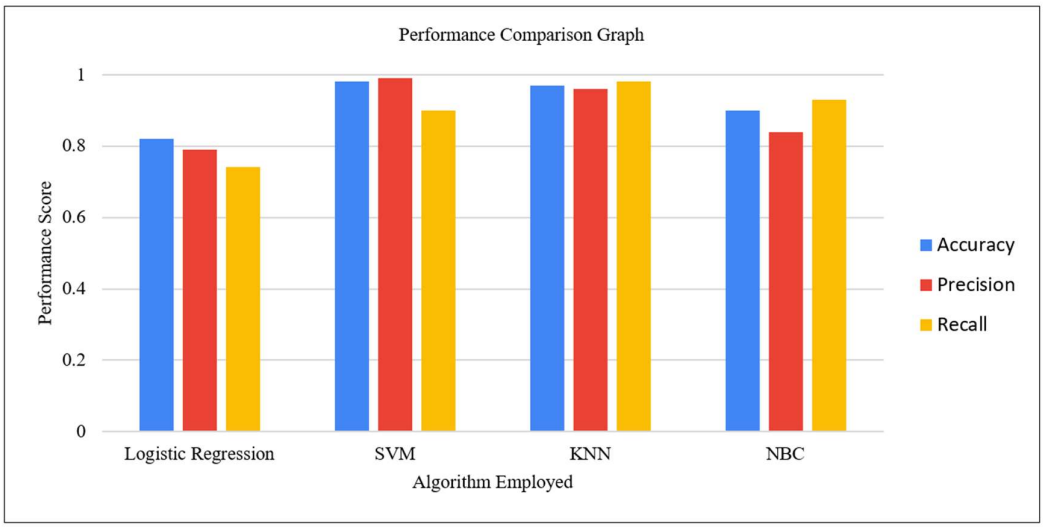

Figure 1. Performance Comparison Graph of all the Employed Algorithms

VII. Conclusion

The goal was to prepare various learning model that could predict the academic performance of a student based on his/her traits in the academic session and then compare the efficiency and performance of these models among each other, to identify which of the trained models would perform the best and provide the most accurate results or predictions in this case. The dataset that was used in this was completely unique and also well suited for the real-world application, as those are the features that any faculty would evaluate and note down at the end of the day. On comparison of all the trained models based on several evaluation metrics, it was found that SVM and KNN would be the most appropriate algorithms or machine learning models for this purpose, providing an accuracy score of about 98% and 97% respectively.

While the development of a student performance prediction model using machine learning algorithms represents a significant step forward in leveraging machine learning for educational purposes, there are several avenues for future research and development:

- Deep Learning: Explore the potential of deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for predicting student performance. Deep learning models have shown remarkable success in various domains and may offer improvements in accuracy and predictive performance.
- Feature Engineering: Investigate additional features that may impact student performance, such as extracurricular activities, parental involvement, and access to resources. Feature engineering plays a crucial role in improving the accuracy and robustness of prediction models.
- Explainable AI: Develop methods for interpreting and explaining the predictions made by machine learning models. Explainable AI techniques enhance the transparency and trustworthiness of prediction models, enabling educators to understand the factors influencing student performance.

- Ethical Considerations: Address ethical considerations related to the use of student performance prediction models, such as privacy, fairness, and bias. Ensure that prediction models are used responsibly and ethically to support student learning and well-being.

By exploring these future avenues, we can further enhance the accuracy, effectiveness, and ethical implications of student performance prediction models, ultimately contributing to improved student outcomes and a more equitable education system.

REFERENCES

[1] G. M. Rao and Prof. K. Kiran Kumar - Students Performance Prediction in Online Courses Using Machine Learning Algorithms (United Int. J. Res. Technol., vol. 02, no. 11) – 2021

[2] C. A. Palacios, José A. Reyes-Suárez, Lorena A. Bearzotti, Víctor Leiva and Carolina Marchant - Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile (Entropy, vol. 23, no. 4, pp. 1–23) – 2021

[3] Harikumar Pallathadka et al. - Classification and Prediction of Student Performance Data Using Various Machine Learning Algorithms (Materials Today: Proceedings Elsevier) – 2021

[4] Leila Ismail, Huned Materwala and Alain Hennebelle - Comparative Analysis of Machine Learning Models for Students' Performance Prediction – March 2021

[5] Nitin Ramrao Yadav and Sonal Sachin Deshmukh - Prediction of Student Performance Using Machine Learning Techniques (ICAMIDA 2022) – 2022

[6] Ziling Chen, Gang Cen, Ying Wei and Zifei Li – Student Performance Prediction Approach Based on Educational Data Mining – IEEE – 23 November 2023

[7] Esmael Ahmed – Student Performance Prediction Using Machine Learning Algorithms – Hindawi Applied Computational Intelligence and Soft Computing – Volume 2024 – 30 April 2024

[8] Hassan g and Udo C. Braendle, "Enhancing Prediction of Student Success: Automated Machine Learning Approach", Article in Computers & Electrical Engineering – 2021.

[9] T. Muncie, "Using Machine Learning Models to Predict Student Retention: Building A Statewide Early Warning System", Morehead State University – 2020.

[10] Ms.Tismy Devasia, Ms.Vinushree T P and Mr.Vinayak Hegde, "Prediction of Students Performance Using Educational Data Mining", ResearchGate – 2020.

[11] Leena H. Alamri, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan and Nida Aslam, "Predicting Student Academic Performance Using Support Vector Machine and Random Forest", ResearchGate – 2020.

[12] Ajibola Oyedeji, Olaolu Folorunsho and Olatilewa Raphael Abolade, "Analysis and Prediction of Student Academic Performance Using Machine Learning", ResearchGate, 2020.

[13] Dr. R Senthil Kumar and Jithin Kumar.K.P, "Analysis of Student Performance Based on Classification and Mapreduce Approach in Bigdata", International Journal of Pure and Applied Mathematics, Volume 118 No. 14, 141–148, 2018.

[14] S. Rovira, E. Puertas, and L. Igual, "Data-driven system to predict academic grades and dropout", PLoS One, vol. 12, no. 2 - Feb. 2017.

[15] Annisa Uswatun Khasanah, Harwati, "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques", IOP Conference Series: Materials Science and Engineering, 2017.

[16] Ihsan A. Abu Amra and Ashraf Y. A. Maghari, "Students Performance Prediction Using KNN AndNaïveBayesian",8th International Conference on Information Technology (ICIT),2017.

[17] Daud,A., Aljohani, N.R., Abbasi, R.A., Miltiadis D. Lytras, Farhat Abbas and Jalal S. Alowibdi, "Predicting Student Performance Using Advanced Learning Analytics", In: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, pp. 415–421 – 2017.

[18] Fiseha Berhanu, "Students Performance Prediction Based on Their Academic Record", International Journal of Computer Application – 2015.

[19] Shanmugarajeshwari, R. Lawrance, "Analysis of Students' Performance Evaluation Using Classification Techniques", IEEE, 2016.